DOCUMENT RESUME

ED 352 973 IR 054 323

AUTHOR Lesk, Michael
TITLE Preservation of New Technology. A Report of the Technology Assessment Advisory Committee to the Commission on Preservation and Access.
INSTITUTION Commission on Preservation and Access, Washington, DC.
PUB DATE Oct 92
NOTE 22p.
AVAILABLE FROM Commission on Preservation and Access, 1400 16th Street, N.W., Suite 740, Washington, DC 20036-2117 ($5 prepaid).
PUB TYPE Information Analyses (070)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Archives; Audiodisks; Audiotape Recordings; Books; Computer Software; *Computer Storage Devices; Copyrights; *Digital Computers; Films; Nonprint Media; *Preservation; Videotape Recordings
IDENTIFIERS Computer Based Archives

ABSTRACT

Intended to stimulate discussion rather than present solutions, this report summarizes how digital technology applies to preservation problems beyond the preservation of print materials. Specifically it describes the problems new kinds of media such as audio and videotape and computer disks pose to librarians and archivists in terms of long-term storage, the varieties of physical format, and the short life of the reading devices. It is suggested that, since the machines to read specific types of tapes and disks become unavailable even before the materials themselves deteriorate, digital preservation depends upon copying rather than the survival of the physical media. New media formats are then discussed both as a problem and as a solution, and the types of formats available in digital form are considered, including audio, video, images of pages, gray scale or color images, text, and numerical data. How to do conversion is then addressed, and issues involved in maintaining computer-based archives are considered. Problems posed by copyright rules are discussed, and four steps that can be taken by librarians and archivists to deal with preservation in the new technological context are suggested. The economic implications of digital preservation efforts are discussed, and the report concludes with a message for librarians in the digital world--(1) preservation means copying and this must be budgeted as a cost; (2) much technical knowledge is involved so cooperation is important; (3) coordinated computer-based archives should be developed, research conducted on conversion technology, and standards developed; and (4) librarians should coordinate with others having the same problems and looking for the same solutions. A list of computer-based archives currently in operation is appended. (ALF)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED352973

The Commission on Preservation and Access

PRESERVATION OF NEW TECHNOLOGY

A report of the Technology Assessment Advisory Committee to the Commission on Preservation and Access

by

Michael Lesk

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

October 1992

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Maxine K. Sitts

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

IR054323

1400 16th Street, N.W., Suite 740, Washington, D.C. 20036-2217 • (202) 939-3400

A private, nonprofit organization acting on behalf of the nation's libraries, archives, and universities to develop and encourage collaborative strategies for preserving and providing access to the accumulated human record.

BEST COPY AVAILABLE

COMMITTEE PREFACE

The Technology Assessment Advisory Committee (TAAC) is a group of seven representatives of industry and academia working in the field of digital technology and its applications in scanning, storage, transmission and printing. The group was charged in 1989 with advising the Commission on applications of electronics for the preservation of and access to deteriorating paper-based materials. This report, one of a series, goes beyond the preservation of print materials. As such, it is a technologist's summary of how digital technology applies to preservation problems. Although authored principally by Michael Lesk, the report represents the views of the entire committee. It has been issued to stimulate discussion, and not to answer all questions.

- Rowland C.W. Brown, Chair
Technology Assessment Advisory Committee

The opinions expressed in this paper are personal opinions of the authors and are not the corporate policy of their employers.

Committee members are: (Chair) Rowland C.W. Brown, Consultant; Douglas van Houweling, Vice Provost for Information Technology, University of Michigan; Michael Lesk, Division Manager, Computer Science Research, Bellcore; Peter Lyman, Librarian and Dean, University of Southern California; M. Stuart Lynn, Vice President, Information Technologies, Cornell University; Robert Spinrad, Vice President, Technology Analysis and Development, Xerox Corporation; and Robert L. Street Vice President for Libraries and Information Resources, Stanford University.

Published by
The Commission on Preservation and Access
1400 16th Street, NW, Suite 740
Washington, DC 20036-2117

October 1992

Reports issued by the Commission on Preservation and Access are intended to stimulate thought and discussion. They do not necessarily reflect the views of Commission members.

Additional copies are available from the above address for $5.00. Orders must be prepaid with checks made payable to "The Commission on Preservation and Access," with payment in U.S. funds.

This paper has been submitted to the ERIC Clearinghouse on Information Resources.

The paper in this publication meets the minimum requirements of the American National Standard for Information Sciences-Permanence of Paper for Printed Library Materials ANSI Z39.48-1984.

COPYRIGHT 1992 by the Commission on Preservation and Access. No part of this publication may be reproduced or transcribed in any form without permission of the publisher. Requests for reproduction for noncommercial purposes, including educational advancement, private study, or research will be granted. Full credit must be given to the author(s) and The Commission on Preservation and Access.

Preservation of New Technology[1]

The rash of new information technology has raised major problems for librarians and archivists, faced with a flood of new kinds of media such as audio and video tape or computer disks. Most of these technologies were designed and manufactured without permanence as a prime consideration. A worse problem is the short life of the reading devices; even the less than permanent tapes or disks often do not deteriorate until long after the machines to read them become unavailable. Digital preservation thus depends upon copying, not on the survival of the physical media. Librarians and archivists must prepare for reformatting as a regular step in information management. In this, they must join with computer center managers, individual computer owners, and all others who need to keep machine-readable data for more than a few years. The good news is that digital media get steadily cheaper and smaller, so that the problems become steadily easier to solve, and that once something is digitized, copying can preserve the information exactly.

New Media Are a Problem

The new media pose several problems to librarians and archivists needing to store them for long periods.

a) For many new media, technological obsolescence is a greater danger than deterioration. Punched cards, for example, were made of relatively strong paper and would last well if kept in proper humidity, but nobody has a reader for them any more.

b) In addition to technological obsolescence, there are also devices that disappear for business reasons: if a particular kind of WORM or magneto-optical disk,

---

[1] It is the policy of Bellcore to avoid any statements of comparative analysis or evaluation of products or vendors. Any mention of products or vendors in this paper is done where necessary to provide an example of a technology for illustrative purposes, and should not be construed as either positive or negative commentary on that product or that vendor. Neither the inclusion of a product or a vendor in this paper, nor the omission of a product or a vendor, should be interpreted as indicating a position or opinion of that product or vendor on the part of the author(s) or of Bellcore. The same applies to the Commission on Preservation and Access, which also does not endorse or evaluate specific commercial products.

for example, is only provided by one vendor, the bankruptcy or commercial distress of that vendor, or even just its change in product line, may render spare parts or replacement readers hard to find.

c) Many of these technologies involve some kind of format, as well as a physical device. Merely the ability to get the bits off the device may not mean they can be easily used.

d) Digital media can be copied without error, in principle; however the equipment required to do so for some kinds of media, e.g., CD-ROMs, is not found in libraries (or even in most service bureaus or other users of CD-ROM).

e) It usually requires some special machinery to look at the information on the object. Thus, mere physical inspection of it will usually not suffice to tell whether it is still in good shape.

f) Format, software and hardware are often intermingled: information may be preserved but if the software to print, search, and edit it has gone, it may be quite costly to make any use of it.

Digital media can be copied without error. *Thus survival of digital information does not depend upon the permanence of a particular object, but upon widespread distribution of the information, and regular refreshing of it onto new technology.* In general the lifetime of the technology is limited by commercial concerns: as technology improves, old storage technologies disappear. For example, half-inch 9-track magnetic tape densities have gone over the last twenty years from 200 bpi to 556, 800, 1600 and 6250 bpi. Today it is difficult even to find an 800 dpi tape drive, let alone a 200 or 556 bpi drive, or any kind of reader for the 7-track format which preceded 9-track on the same kind of physical tape. And now all half-inch tape equipment is threatened by the more compact DAT (digital audio tape) and 8mm videotape devices.

From the standpoint of a preservationist this example is a particular problem because the new materials are less durable (both DAT and 8mm video are helical-scan tapes, which generally have inferior lifetimes as compared to linear recording tapes). The new systems are changing too: we already have two styles of 8mm videotape. *Thus, the only security is copying to new devices regularly.*

Reformatting, instead of being a last resort as material physically collapses, will be a common way of life in the digital age. The steady decrease in the cost and size of digital media would make this attractive even if we were not driven to it by technological obsolescence. As new media become widely used in the computer industry, and old media become ever less available, libraries should expect to transfer their digital holdings, either by transcribing files themselves, or (more likely) by buying copies on new media from

somebody else. These new copies should be not only in a newer physical format, but in whatever software format is closest to standard at the later time. Peter Lyman, University Librarian and Dean, University of Southern California, points out that while paper deteriorates when used, digital media often not only do not suffer from being used, but benefit from use: if something is used, the librarian will find out faster when the users want it on a new device.

Surprisingly to the novice, the software formats are often longer-lived than the physical devices. Unix word processing software from 1975, for example, is still in use; the storage devices (low density magnetic tape, removable disk packs, and so on) of that time are gone. Unfortunately there is also a much wider variety of logical formats, and much more varied expertise is required to deal with the software content than with the physical material. The same is true traditionally, of course: a specialist in paper conservation can work on a book regardless of the language it is written in, while a specialist in cataloging one discipline or reading one language normally works only with books meeting that condition. Similarly, it takes fewer experts to keep up with tape and disk machines than it does to keep up with the much wider variety of word processing systems.

On the good side, the intervention of machinery between the actual object and the reader means that the users are unlikely to become emotionally attached to the particular physical media, and thus reformatting of advanced technology should not produce the objections that accompany reformatting of books. It is not expected that computer media will have any artifactual value, except odd examples in the Smithsonian. Also, the steady progress of technology means that the difficulty of keeping any particular item decreases with time. To return to the discussion of punched cards, nobody would want to keep them in that format any more, even if it were practical, since at only 80 bytes per card, 100 pounds of punched cards fit onto one 3.5 inch 3/4-ounce diskette today.

New Media Formats Are a Problem

What physical formats are involved? There are many kinds of new format technologies, including:

a) current analog technologies: audiotape cassettes, photographs, movie film, and VHS videotape.

b) obsolete or semi-obsolete analog technologies: phonograph records (33, 45 and 78 rpm), Beta videotape, 8-track audio cassettes, and reel-to-reel audiotape.

c) obsolete portable microcomputer storage devices: 8-inch floppy disks, various adaptors for conventional audiotape cassettes, and, one can predict, 5.25 inch floppy diskettes.

d) current microcomputer storage devices: 3.5-inch rigid floppy disks (an oxymoron of a name, but common), QIC-type cartridges in various lengths and densities, DAT (digital audio tape), and CD-ROM.

e) obsolete larger computer devices: punched cards, removable disk packs of various kinds, various kinds of magnetic tape including Dectape, 7-track half-inch tape, 9-track tape in densities below 1600 bpi, punched paper tape, and numerous other devices made briefly (I still have a reel of steel tape used by the Univac I).

f) current larger computer devices: Exabyte 8mm video cartridges used as digital media, DAT cartridges, 9-track tape in high densities of which 1600 and 6250 bpi are most common, magneto-optical disks in ever-increasing sizes and densities, and WORM disks also in increasing sizes and densities.

These devices vary from some which are quite durable (SONY claims a 100-year life for its large WORM cartridges) to others which are fragile (5.25 inch floppy diskettes were often very cheaply made and will not stand hard handling). For those interested in the physical durability of the new media, there is a good review by Saffady.[2] However, it is not usually durability, but rather obsolescence, that poses a threat to the material on the new technologies.

Analog media pose even worse problems. In addition to the technological obsolescence shared with digital media, and which is now wiping out the vinyl record and the devices to play it, for example, the analog media cannot be copied without deterioration and are not necessarily very durable. Both audiotape and videotape, for example, are comparatively fragile, especially if used; a book can be read many more times without falling apart than a videotape can be played without major deterioration.

Since analog material degrades with copying, it should be converted to digital at the first practical opportunity, and carried forward in digital form. Digital material is not only copyable without error, but it gets steadily cheaper to store with time. The sooner the librarian converts to digital, thus, the better off we are. Against this is the high cost of conversion, and there is a danger that a low-quality conversion will cause many users to wish to go back to the originals. For example, Michael Ester of The Getty Conservation Institute has measured the sensitivity of art historians to different quality digitizations of the same photograph, and demonstrated the important of good quality. Unfortunately, business practices in this area are counter to what libraries would like: some businesses digitize photos only at low resolution not only to save money, but to be sure that no one can use the digitized version (relatively easy to copy) as a substitute for buying the original (somewhat harder to copy illegally).

---

[2] William Saffady, "Stability, Care and Handling of Microforms, Magnetic Media and Optical Disks," *Library Technology Reports*, vol. 27, p. 5-117, (January-February 1991).

Converting to digital is now feasible for sound, but digitizing video is still not a routine process which libraries can use. There are also some who think that even with sound, the digital copy is degraded from the analog version (at least enough to keep several hi-fi magazines in business).

New Media Are the Solution

Digital storage is made of bits. The storage and preservation of these bits is independent of what they represent. The interpretation of them, on the other hand, is independent of the medium on which they are stored. Thus, conversion and storage break down into two parts:

i) on what physical device should the bits be stored?
ii) in what format should they be stored, and how will they be interpreted?

The physical devices, as mentioned, change regularly. Any solution should be viewed as temporary. A library may, however, choose a more or less temporary solution. Compare, for example, WORM optical disk platters with 8mm or 4mm helical-scan tape. The following table shows some basic parameters (SONY 12-inch WORM platters are chosen as an example):

Properties of Some Digital Media (1992)

	WORM	8mm Exabyte	DAT
Capacity (Gbyte)	6	5	1.3
Cost	$340	$5	$9
Wt (oz)	40	3	1.5
Dimensions (inches)	12.75x13.5x.6	4.0x2.6x.75	3x2.6x.6
Volume (cu in)	103	5.5	2.4
Long life?	yes	no	no
GBytes/cu ft	100	1100	480
GBytes/lb	2.4	27	14
MBytes/$	17	1000	140

Of course, the money spent on the WORM platter gives many benefits. Its life is expected at 100 years, while the tapes are probably only good for a few years. The platter is much more resistant to temperature and humidity variation. It is less likely to be destroyed by a faulty mechanism. But the most important advantage for system designers is the random-access property of the platter, and this makes it more attractive for online use, rather than for archiving. If the only question is preservation, making and distributing multiple copies of the cheaper media is more sensible and protects against other possible problems (e.g., fire or theft). The row "long life" does not mean permanent enough to keep forever -- it only reflects the relative lifetime of the devices here, all of which may turn out to have physical lives that exceed their technological life. The library with only a small number of

tapes or platters, of course, may find that the cost of the reader also matters -- a reader for the SONY WORM platters costs about $20,000 while DAT and Exabyte drives cost only $1000-$3000 today. This table is somewhat biased in favor of the physically smaller media, since it ignores the need to arrange them in some way that permits access; one cannot fill a cubic foot with DAT cartridges and expect to find one that is in the middle.

If the material must be stored on-line, with relatively close to random access, the space and costs are considerably higher. Here are some approximate numbers.

Costs of On-Line Storage (1992)

	WORM	Magneto-Optical	Winchester	CD-ROM
	jukebox	jukebox	magnetic disks	
Capacity(GB)	300	40	2.2	.55
Cost	$150K	$30K	$3.7K	$.5K
Size (cu ft)	20	4	.1	.3
GB/cu ft	15	10	20	2
MB/$	2	1.3	.6	1

Not surprisingly, these costs are much higher than those for the demountable storage devices. But again, with these media the time required to present any random item is measured in either seconds (for the jukeboxes) or milliseconds (for permanently mounted media), whereas to read through a full Exabyte cartridge may take an hour. Transfer rates are also generally higher on the disk devices. If the primary intent of a library is access, rather than only preservation, the amount of delay and the need for an operator to mount a tape or disk must be weighed against the cost.

The above tables are only true in 1992. For example, ICI has developed a material they call "digital paper" which offers a promise of even higher density and lower cost, along with long life and non-erasability. It is too soon, however, to say whether this is the likely next device on the market. Currently, the biggest activity is in rewriteable 3.5 inch magneto-optical disks, but the capacities are fairly small in our terms (128 Mbytes).

Unfortunately, just having the bits on some media is not enough. The user needs to be able to find and interpret the bits he or she wants. In some cases, vendors use proprietary software to gain access to data and do not provide details, sometimes as a way to prevent illegal copying. Such techniques further aggravate problems of future access, and require attention to the availability of the software that reads a particular vendor's data, as well as the availability of the device. Standards may help in the future.

What Formats Ought We to Have?

The question of format is much more complex. A great many kinds of information are kept in digital forms. No simple table can summarize the choices. Most of the decisions will be made far outside the library and archives community, and because of the economies and practicalities of sticking with what is decided for general industry, libraries and archives should expect to go along.

1) ***Audio.*** Two formats are likely to be common. One is the CD standard, a high-quality 176 Kbytes/second stereo signal. The other is ISDN telephone quality, 8K bytes/second. For music the first is necessary -- the telephone quality standard chops frequencies above 4 Khz and provides only one channel. Undoubtedly the committees that will consider this point will recommend that CD audio should be used for everything. Personally I think that the 8K signal is quite adequate for speech. There are even more compact signals possible for speech, but there is not enough need for further compression to justify their use, and none seems to be winning general acceptance. Converting audio to the 8K or 176K format is practical today and makes sense for oral history collections (which would actually take less space, if that mattered, stored as 8K digital audio than as ordinary cassettes).

2) ***Video.*** Digital video at present is in D2 format, a signal with a great many bits in it (165.888 Mbit/sec), but limited image quality. In particular, even 16mm film cannot be converted to this format without noticeable quality loss. D2 equipment and tapes are also quite expensive (this format is designed for and used by professional broadcast stations). Two new standards are likely: whatever comes for HDTV (high definition television) and whatever is decided for MPEG (Motion Picture Experts Group, a new compression algorithm). HDTV will be a high quality standard, but international agreement is probably a few years away. The MPEG standard will be finished soon, but represents a lower image quality (although acceptable for material now on VHS NTSC videotape). Expect MPEG to be in the neighborhood of 256Kbytes/second (that's about 1G per hour) and HDTV to be perhaps 8 times more bulky. Conversion is not practical for libraries, as much of the equipment is very expensive. Remember that this is a different league -- the entire budget of the Harvard library system for a year is less than the cost of making one blockbuster movie. Also note that MPEG is a lossy compression algorithm not designed to preserve its input bit for bit, but to provide the same effect on the viewer.

3) ***Images of pages.*** Most printed pages are high contrast (not containing halftones or color) and a one-bit-per-pixel representation makes sense. Group IV fax seems to be the best compression scheme in common use today. It is a lossless algorithm: the decompressed image will be exactly the same as the

original. There is an additional issue of file packaging surrounding the actual image data; TIFF (Tag Independent File Format), in some variant is the most acceptable today. Alternate forms in which these images might arrive include the screen dump formats of various window systems and manufacturers, and the formats used by a variety of programs such as MacPaint. Conversion software exists for most of these; for example, the public domain package "pbmplus" accepts among other formats, CMU window manager, FITS, Facesaver, fax, GEM ".img" format, HIPS, Sun icons and rasters, Amiga IFF ILBM, Macpaint, MGR, X10 and X11 bitmaps, PC paintbrush and doodle-brush, Postscript, and X window dumps.

If a library is doing its own scanning, resolutions range from 200 dpi (fax quality) to 300 dpi (laser printer quality) to 600 dpi. Three hundred dpi is good enough for readability down to 5 or 6 point type, assuming that there are no halftones on the page. However, there are many pages for which more careful treatment is desirable: (a) pages for which quality reproduction rather than just readability matters, (b) pages containing halftone images, which are likely to be seriously degraded by the typical 300 dpi one-bit-per-pixel scanner, in the same way that an ordinary photocopier would degrade them, or (c) pages which are handwritten, faint, or otherwise requiring special attention. Higher resolution is not the entire answer: in addition to high resolution, grey-level scanning may be needed. Fortunately scanning equipment continues to improve: 900 dpi scanners will be practical before long. In terms of compression, in Group IV fax it will be found that the size of the image varies roughly linearly with the resolution, rather than quadratically as in raw scanning data. Expect to need 50K-100K bytes per large densely-printed page at 300 dpi. Conversion is practical today at medium densities.

4) ***Grey scale or color images (e.g., photographs).*** In principle the answer to this is the JPEG (Joint Photographic Experts Group) standard, which is almost but not quite practical. The amount of storage needed will depend upon the image quality demanded and the size of the image; expect 200K-750K bytes or so for a high quality representation of a 6x9 or so image. Note that normal JPEG is, like MPEG, a lossy compression algorithm: the bits you get back may not be exactly what you put in. Storage efficiency is improved by making slight changes to the image. Depending on the quality and the source of the original image, and also on the intent of the system, a different algorithm (e.g., JBIG or TIFF) might be preferable, or JPEG used in a lossless mode

5) ***Text.*** For most library and archives purposes this is likely to be the most important immediate issue. What will arrive will be a variety of word processor formats -- Word, Wordperfect, Macwrite, Nota Bene, Troff, LATEX, FrameMaker, Interleaf, Multimate, Scribe, and regrettably many others plus

variations within them (such as the macro packages common in Troff). Typesetting tapes from various proprietary printing systems are also going to show up at archives. There are various groups working on attempts at standardization, including AAP (Association of American Publishers) with its Electronic Manuscript Standard, and the Text Encoding Initiative mentioned before.

Unfortunately there is no general conversion package, and in principle there cannot be. Many of the word processing documents will arrive with typographic markup -- i.e., commands such as "italic" or "bold" in the document. The AAP and TEI standards want intellectual markup -- i.e., instructions such as "journal title" or "section heading." There is no general and simple way to deduce the intellectual markup from the typographic markup. For straightforward typographic transcription of most of the word processing formats, commercial products are available today (e.g., Softswitch). But the insertion of true intellectual markup is a task comparable to copy-editing of a text; it adds significant value to any future use, but represents a lot of work.

6) ***Numerical Data.*** This is another important immediate issue for which, unfortunately, standardization is much further off. Different standards groups are working in different subject areas, and there is little hope of any common representation that will cover everything from census data to astronomical tables. In some subjects, the input format of common spreadsheet programs such as Lotus-1-2-3 is common, but we do not have an answer to the problem of self-identifying data that can be read and used without first studying a lot of documentation. Librarians used to the arguments about LCMARC, UKMARC, and EuroMARC will understand some of the problems and the frustrations. There are a great many areas in which numerical data interchange is well developed, e.g., protein structures (Brookhaven and the European Molecular Biology Laboratory hold files), the gene sequencing data banks, satellite imaging, and so on. Individual laboratories such as the National Library of Medicine do a great deal of research on the exchange of particular forms of numerical data, and libraries and archives need to be aware of this work and its implications for long-term storage.

On balance, libraries and archives can focus for the moment on text and audio. These are practical, and there is some understanding of what should be stored. Images are just about ready for the more advanced organizations to try reformatting. For data and video, libraries and archives can probably stage holding actions with current technology until progress is made by outside standards committees.

How To Do Conversion

Technology for doing conversion of older digital formats to newer formats, either physical or logical, is still quite primitive. The text formatting situation is particularly serious. When faced with a new typesetting tape, for example, it is still sometimes necessary to resort to an octal dump followed by comparison with the printed book. The documentation for the internal formats of many systems is inadequate, either for proprietary reasons or simply because it is not worth doing well if the typesetting system is not widely distributed.

Even if the previous format is known, the converter is normally reduced to writing a specialized program. Sometimes languages like Lex will help, but often there is no really appropriate tool. There has been some work on software to assist in format conversion, for example the Chameleon project at Ohio State, but nothing that is generally accepted.[3] As mentioned before, the problem is that most typesetting and WYSIWYG formatting systems deal only in appearance (e.g., italic, bold) and not in meaning (e.g., cited journal name, section heading). It may or may not be possible to guess at the correct meaning in a particular context, but there are no programs that do so in general.

The situation is somewhat better for graphics where the problem is purely one of format and not of intellectual content. Among the public domain graphics packages are:

Package	Author	FTP address
PBMPLUS	Jeff Poskanzer	ftp.uu.net:/networking/osi/quipu/pbmplus.tar.Z
Fuzzy Pixmap	Michael Mauldin	ftp.uu.net:/graphics/fbm.tar.Z
Img-whatnot	Paul Raveling	ftp.uu.net:/graphics/urt/urt-img.tar.Z-split/img_1.3.tar.Z
Xim	Philip R. Thompson	world.std.com:/scr/X11R4/xim3g.tar.Z
TIFF Library	Sam Leffler	ftp.uu.net:/graphics/tiff/v3.0.tar.Z

In addition, various commercial packages address the same problems.

The situation for data is even worse than for text; so far, special documentation is required for each file. The only remotely agreed-upon structure is relational files, i.e., records with fields, and there is no standard way to write relational databases to an interchange format nor is there likely to be any. The point, of course, is not just to get the numbers out but to have them identified in some way.

Scanning paper is perhaps the one part of conversion which is fairly well understood: the Cornell project on preservation of books by scanning and reprinting, for example, is finding costs of about 10 cents per page to scan a book of 300 pages, plus storage and refreshing costs of under 1 cent per page per year (again for books of 300 pages).

---

[3] S. A Mamrak, M.S. Kaelbling, C.K. Nicholas and M. Share, "Chameleon: a system for solving the data-translation problem," *IEEE Transactions on Software Engineering*, vol. 15, no. 9 pp. 1090-1098 (1989).

Computer-based Archives

There is great economy of scale in maintaining new media -- not so much the devices, but in sharing the expertise. Centralized computer-based archives exist today -- for example at Oxford University for literary texts, or the University of Michigan for social science data. Cooperation is essential for libraries to manage. Not only will standard formats help a great deal, but copying digital information is much easier than generating it. Particularly where lots of intellectual effort is required (e.g., indexing pictures or audio) it makes no sense for different groups to work on the same file. It will be much easier for libraries to obtain updated versions from central sources than for each library to manage all its own reformatting. A cooperative program for converting digital information would be appropriate. It should involve both maintaining records of who has what, seeing that conversions are done only once, and seeing to it that important materials are kept in more than one place.

There are a great many existing computer-based archives. In the Appendix is a short list, several years old, of machine-readable archives within Canada, the UK and the US which was collected by Prof. Robert Kraft of the University of Pennsylvania. It is shown merely as an example of the kinds of centralized computer-based archives which already exist; lists can be found both online and in conventional publications.[4] There are already secondary services which collect knowledge about machine-readable texts, such as the Center for Electronic Texts in the Humanities (Rutgers and Princeton), whose catalog is in turn online under RLIN.

Efforts should be made to find funding to put archives of this sort on a firm footing. In the long run, it will be cheaper to have sites of responsibility for particular materials than to attempt to duplicate expertise in specialized records conversion. Furthermore, sites of this form will function in the traditional way to link scholars working in the same area, even if these links are now via electronic mail rather than by physical travel to the site.

Most of the computer-based archives above are textually oriented. Support is needed to extend computer-based archives in the multimedia areas, and also in the areas of numerical data and programs. Some attention is also needed to ways of indexing the material in these archives, as there is no straightforward way to locate information, particularly non-textual information, in machine-readable form. Bibliographic guides will have to be developed, and even the format these should take is not clear.

Industrial and government computing groups have often had the problem of maintaining old tape files, and have stimulated much of the work in records management for machine-readable files. Unfortunately the history here is often poor: it is said that we know

---

[4] For example, see "The Very Pulse of the Machine: Three Trends Toward Improvement in Electronic Versions of Humanities Texts," by Michael Neuman, *Computers and the Humanities*, v. 25, p. 363-375 (1991); and "The Humanities Computing Yearbook, 1989-1990," edited by Ian Lancashire and Willard McCarty, Oxford University Press, 1991.

more about the 1860 census than the 1960 census, which was written on a very early form of computer tape and which no one realized in time should be copied to newer media.

Copyright Is Always a Problem

Copyright rules may pose a severe problem. The copyright life of much material greatly exceeds the technological life of the media it is recorded on. A library may well be able to preserve something only by copying it, but may have difficulty obtaining permission to do that. With luck the original issuer will publish the material in a new format, but as the devotees of old records and movies can testify, this does not always happen. Leaving aside the question of how it is accomplished, through either amendment or court interpretation, the concept of "fair use" should be extended to include the copying to new media of a publication which is no longer available for sale, and the distribution of the copy to other libraries. At the moment, there is an exemption for copying of a publication that a library already owns, but not for distribution to libraries that do not own the original being copied. In the next century, as the United States implements a "life plus 50" standard for copyright expiration, it may even be hard to determine whether or not a work is still copyrighted.

The law may also pose problems even for information where the copyright law does not bar conversion. A library or archives might wish to store information derived from old, copyright-expired publications; or from public domain sources; or from publications of the university to which it belongs. However, it is doubtful that a copyright can be asserted in a straightforward transcription to machine-readable form of a non-copyrighted document or audiotape. The rule that a copyright can only be claimed for the result of creative activity is somewhat unclear. It has recently been interpreted to say that (a) collecting and alphabetizing names into a telephone book is not copyrightable[5] but at the very least (b) numbering the pages in a Government-provided text is copyrightable.[6] Although measuring the creativity in these publications is like arguing whether a concrete block is more or less intelligent than a 2x4, the legal uncertainty causes all vendors of electronic information to require specific contracts limiting the use of the purchased data, imposing administrative burdens on everyone.

Libraries and archives, depending on the circumstances, might be on either side of this issue. Presumably, the ability to enforce a copyright in a converted text will encourage someone to do a conversion. On the other hand, if the issuer then disappears, it becomes difficult to get permission to do the next conversion when that becomes necessary. Again, it would be useful to have a rule that the copyright lasts only as long as someone is willing to sell the protected item.

---

[5] *Feist Publications, Inc. v. Rural Telephone Service Co., 111 S. Ct. 1282 (1991).*

[6] *West Publishing Co. v. Mead Data Central, Inc., 779 F.2nd 1219 (1986).*

The copyright law also affects the question of multiple copies. The United States is developing a very high speed digital network which will soon provide 45 Mbits/second between major sites and should be at 1 Gbit/second before the end of the decade. Thus, libraries and archives may well feel that there is no need to have their own copy of something, since the delay in getting it remotely may be negligible. But, depending on copyright rules and the developing business practices in the electronic information industry, it may be difficult to arrange access permission. On the other side, it would be unfortunate if the existence of the network meant that important documents were kept in only one copy which was at risk of disappearing (e.g., residing in some individual's personal disk files, rather than being kept in some permanent institution's records management system). Finally, the insertion of copy-protection into sound recordings acts against the interests of libraries and long-term permanence of the original material. Deposit of an unprotected copy in some trustworthy institution is probably the best solution, but not something the industry is pursuing.

Recommendations

In this new world, preservation means copying, not physical preservation. Librarians and archivists, already involved in this basic idea through efforts in reformatting, can cooperate in learning how to deal with preservation in the new technological context. What steps can be taken?

(a) Computer-based archives can be encouraged, along with the cataloging and central deposit of machine-readable data.

(b) Where appropriate, software standards can be encouraged. It is often more difficult to salvage an old format than an old data medium. Which of us still has a formatter for the 1960s "runoff" program, or a translator for the IBM 7094 BCD character set? Among the proposed standard formats today are SGML (standard generalized markup language), especially as specified by the Association of American Publishers in its "Electronic Manuscript Standard", or ODA (open document architecture). Regrettably, it is often the case that standards still leave a great deal of choice. Simple one-bit-per-pixel bitmaps, for example, can come in a dazzling variety of formats (some of which were listed earlier). Groups which are attempting to help simplify the variety of formats, such as the Text Encoding Initiative for literary material, can be encouraged.

(c) Remembering, however, that two hundred years after the invention of the metric system it is still not universally accepted, we must accept that conversion will always be with us. It would be most productive to encourage research projects in conversion methodology: software tools for changing from one format into another. This might include research on fast emulators and simulators, on translation methods, and related areas.

(d) Computer museum operators can be urged to maintain software as well as hardware, and to be able to operate old programs for purposes of translation. This software will have to be run through emulation -- the spare parts problems will make it impossible to keep old machines running.[7]

Most important, librarians and archivists can adopt a new view of preservation with digital media. It means copying, not physical preservation, and it requires more attention to long-term costs. Some analog media are also best preserved by copying, of course. There are analogies to the old problems in the new era, but not exact ones. For example, rebinding is irrelevant with digital media, where no one is attached to the physical form of the device; instead we have refreshing, or copying. Theft is still a problem, but it is more likely to be in the form of illegal copying rather than stealing a physical device. Mis-shelving of books is gone, but computer viruses and accidental or deliberate erasing (e.g., to gain an advantage in a competitive undergraduate course) will replace it. Systematic problems such as acidic paper correspond to the unfortunately much more rapid bankruptcy of supplies. Not everyone will warm to the analogy of buying new clothes because the styles have changed rather than because the old ones have worn out, but that's the way computer devices develop.

The preservation task for new materials does share with microfilming a need for collaboration. A centralized clearinghouse of some sort is needed to avoid doing the same conversion twice; and it is likely that every library will not wish to be involved in creating digital material in new formats, most merely using it, and creating it through consortia or contract houses.

In particular, cooperation on these issues is needed not only among librarians and archivists, but with computer centers and others involved in records management. The problems facing the libraries are also faced by every computer operator, and efforts should be made to share information between all affected groups.

Of the technical issues, the most important is to develop better conversion technology. It is extremely frustrating today to encounter unfamiliar text in machine-readable form: often considerable human expertise is needed to extract what the user wants from it. Research in ways of converting formats or adapting old software would be particularly important.

Economic Implications

What will this cost? Let us try to imagine the librarian of a few years from now buying a book. First, consider a model just like the library of today -- shelves full of stuff, readers sitting at tables -- except that the shelves are full of tape cartridges and there are computer screens on the tables in front of the readers.

---

[7] My thanks to Professor David Gifford of MIT for raising this issue.

Suppose that a book is purchased as a batch of images on tape from the vendor; assume 300 pages at 100 Kbytes per page, or 30 Mbytes total. Perhaps it is purchased on a DAT cartridge; 30 books will fit on the one cartridge, whose material cost might be $10. This sounds like the book will only cost 35 cents, but remember that the paper, ink, cloth, glue and postage for a $25 book does not cost much more than $1 today. So the book will probably still cost $25; the publisher still has all the other costs.

However, every few years the library will have to buy a new device and copy the tape. Copying such a tape takes about 30 minutes, or 1 minute per book. The major part of the cost will be the new tape (suppose another 70 cents) and the cost of somebody to watch the drive (at $10/hr, perhaps 20 cents). Thus, a superficial estimate suggests that it will cost $1 every 3 years, or 30 cents per year, to refresh the book. In addition to the $20 to buy the book, therefore, the library would have to set aside enough money to generate 30 cents per year, or perhaps another $6. More thorough estimates at Cornell University suggest a cost of $1.37 per book per year for refreshing.

This may sound like a disadvantage; but remember that the single cartridge which holds 30 books fits in half the space of an ordinary book. Assume that shelf space normally costs in the neighborhood of $1 per year per book,[8] and that half of that is the kind of building cost which varies proportionally to square feet; then, this is reduced from $0.50 to $0.01 for the cartridges. This saves almost 50 cents per year per book, and will compensate for much or all of the cost of copying, provided that the accounting system of the university or other organization to which the library belongs can actually equate the capital cost of providing shelf space to the running cost of performing digital copies.

If the books can be bought as typesetting tapes instead of images they will occupy 1/20 or so the digital space, and thus the total costs of the computer media and operations will decline, leaving the library somewhat better off. Most importantly, it is more likely that a library will be able to bulk purchase a large number of books from some archive or publisher cooperative at a reasonable cost, and expect them to maintain the material as formats change.

It might be thought that libraries will be able to get revised formats of books from the original publisher. However, it is not clear how the pricing will work. At least today, no publisher sells a second edition of a book to a library at a reduced cost if it owns the first edition. Furthermore, the vast majority of books come out in only one edition, and it is not clear that will change.

---

[8] Hayes, Robert M. "The Magnitude, Costs, and Benefits of the Preservation of Brittle Books: Reports on the Preservation Project." Washington, DC: Council on Library Resources, November 30, 1987. To justify this premise, consider that in a square foot of floor space a library can put about 80 books -- 1 book per inch, shelves about 12 inches deep with support, and six to seven shelves high. Assuming that the shelving requires a square foot of aisles for each square foot of shelves, this means about 40 books per square foot. Renting commercial space with heat, power, etc. is about $40 per square foot per year.

With time, this process will at least be getting cheaper. As storage densities increase, whatever information is being copied today will fit in less space tomorrow (for example, a single Exabyte cartridge will hold the equivalent of 95 reels of 800-bpi magnetic tape, and the cartridge will fit in 6 cubic inches while the 95 reels of tape need more than 6000 cubic inches).

The other obvious model is that the library does not buy most books, but obtains them on-demand from some centralized archive. In this model, the library still contains the tables, and has some machines on them, but the book content is not obtained from any kind of local stack, but from a remote link to a centralized computer-based archive. Users do searches in secondary services or in catalogs, and then retrieve papers or books from the archive servers. What will the data transmission cost? At the moment, most Internet sites pay a bulk once-a-year price; there is no charge for individual messages. What might a charge be? At present they are still rather high: for example, A DS-1 (1.5 Mbit) line across the U. S. costs about $600 per hour. That means that $1 will move about 1.2 Mbytes, so a 30 MByte book would cost $25. But the NREN is likely to be much cheaper, at least if present political and technical trends continue. Just as today, postage is not a major part of the cost of getting books onto the shelves, in the future, data transmission is not going to be most of the cost of getting bits into the library. In fact, the costs of data transmission are likely to be much less than the savings of not having to locally catalog and manage the books.

The major issue will be copyright permission. It may be technically feasible to move bits even today, but administratively it may be a horror. Work is needed to sort out the legal and practical problems of making file-sharing possible. Some researchers (e.g., Brian Kahin of Harvard, John Garrett of Corporation for National Research Initiatives, and Marvin Sirbu of Carnegie Mellon University) have already been working in this area. What is needed most are some prototype projects, to find suitable business models and see which of the problems envisaged are real and which imaginary, and sketch out a suitable balance by which the users, libraries, publishers/authors, and information industry vendors share in the advantages and the costs of the new technology as applied to books. The computer software industry, for example, seems to be converging on site licenses or concurrent-use licenses as a reasonable pricing mechanism; perhaps this will serve as a guide to the libraries.

Conclusions

The message for the librarian in the digital world is that

(a) Preservation means copying; this must be expected from the beginning and budgeted as a cost;

(b) There is a lot of technical knowledge involved, so cooperation is important;

(c) Various groups, including the Commission on Preservation and Access, should urge development of coordinated computer-based archives, research on conversion technology, and standards activities.

(d) Librarians should coordinate with computing center operators and records managers in other areas who have the same problems, and with the publishers and vendors who are looking for the same solutions.

Acknowledgments

The comments of Clifford Lynch, Stuart Lynn, and Peter Lyman have been very helpful in preparing this article.

Appendix I – Examples of Archives

McGill (Univ): Kierkegaard-Wittgenstein Project
Montreal (Univ): Instut d'Études Médiévales
Montreal (Univ Quebec): Centre d'Analyse de Textes par Ordinateur
Newfoundland (Univ): Folklore & Language Archive
Ottawa (Dept Communications): CHIN = Canadian Heritage Information Network
Ottawa (Carleton Univ): Centre for Editing Early Canadian Texts
Quebec (Laval Univ): Bibliographical Information Bank in Patristics
Toronto (Univ): RIM = Royal Inscriptions of Mesopotamia Project
Toronto (Univ): REED = Records of Early English Drama
Toronto (Univ): CCH = Centre for Computing in the Humanities
Toronto (Univ): Dictionary of Canadian Biography
Toronto (Univ): DOE = Dictionary of Old English
Toronto (Univ): DEEDS Project = Documents of Essex, England Data Set
Toronto (Univ): Greek Index Project (Pontifical Institute of Medieval Studies)
Waterloo (Univ): Centre for the New Oxford English Dictionary
Vancouver (Univ British Columbia): UBC Data Library
Vancouver (Simon Fraser Univ): RDL = SFU Research Data Library

Cambridge (Univ): LCC = Linguistic Computing Centre
Edinburgh (Univ): EUDL = Edinburgh University Data Library
Edinburgh (Univ): Greek Text Database
Essex (Univ): ESRC = Economic and Social Research Council Data Archive
Glasgow (Univ): DISH = Design and Implementation of Software in History Project
Lancaster (Univ): UCREL = Unit for Computer Research on the English Language
Leeds (Univ): Centre for Computer Analysis of Language and Speech
London (Univ): School of Oriental and African Studies
London (Univ College): Survey of English Usage
Oxford (Press): OED = Oxford English Dictionary
Oxford (Press): Oxford Shakespeare
Oxford (Univ): OTA = Oxford Text Archive
Oxford (Univ): Lexicon of Greek Personal Names (Bodleian Library)
Southampton (Univ): AIE = Archaeological Information Exchange
York (Univ): Graveyard Database

AZ Tucson (Museum): Documentary Relations of the Southwest
CA Berkeley (Univ CA): Anthologies of Italian Music and Lyric Poetry of the Renaissance
CA Berkeley (State??): SDB = State Data Bank [supplements ICPSR/Michigan]
CA Berkeley (Univ CA): Sino-Tibetan Etymological Dictionary and Thesaurus
CA Davis (Univ CA): Project Rhetor
CA Irvine (Univ CA): TLG = Thesaurus Linguae Graecae
CA Los Altos (Inst): PHI = Packard Humanities Institute
CA Los Angeles (UCLA): Computerization of Arabic Biographical Dictionaries for the Onomasticon Arabicum
CA Malibu (Udena Publ): CAM = Computer-Aided Analysis of Mesopatamian Materials
CA Menlo Park (Cent): Center for Computer Assisted Research in the Humanities
CA Riverside (Univ CA): Laboratory for Historical Research
CA Riverside (Univ CA): Biographical Data Base for the Soviet Bureaucracy
CA San Diego (Univ CA): International Electronic Archive of the Romancero
CA Santa Barbara (Univ CA): Domesday Book Database
CA Santa Monica (Getty Art Hist Info Prog): Provenance Index
CA Stanford (Univ): Institute of Basic German
CO Colorado (Univ): Siouan Languages Archive
CO Boulder (Univ CO): CCRH = Center for Computer Research in Humanities
CT Hamden (??): Encyclopedic Thematic Catalog of Russian Sacred Choral Music
DC Washington (Georgetown Univ): Electronic Text Repository
DE Newark (Univ DE): Massachusetts Tax Valuation List of 1771
FL Talahasee (FL State Univ): Center for Music Research
HI Manoa (Univ HI): Salish Lexicography
IL Chicago (Univ): ARTFL = American and French Research on the Treasury of the French Language
IL Chicago (Newberry Lib): County Boundaries of Selected United States Territories/States
IL DeKalb (N IL Univ): Tai Dam Dictionary and Text on Computer
IL Urbana (Univ IL): Hymn Tune Index
MA Boston (Center??): Census of Gothic Sculpture in America

BEST COPY AVAILABLE

MA Cambridge (Harvard Univ): Boston Dainas Project
MA Cambridge (Harvard Univ, Boston Univ): Perseus Project
MA Williamstown (Getty Art Hist Info Prog): Art and Architecture Thesaurus
MD Baltimore (Johns Hopkins Univ): CAL = Comprehensive Aramaic Lexicon
MI AnnArbor (Univ MI): Family Life and Conditions in the US, 1888-1936
MI AnnArbor (Univ MI): ICPSR = Inter-university Consortium for Political and Social Research
MI Dearborn (Univ MI): Comprehensive Computer Data Bank of the Medicinal Plants of Native America
MS Hattiesburg (S MS Univ): Faulkner Computer Concordance
NC Chapel Hill (UNC): DBAGI = Data Bank for Ancient Greek Inscriptions from Athens
NC Durham (Duke Univ): DHDB = Duke Humanities Data Base
NC Durham (Duke Univ): DDBDP = Duke Data Bank of Documentary Papyri
NC Winston-Salem (Museum): Index of Early Southern Artists and Artisans
NH Hanover (Dartmouth Univ, Princeton Univ): Dartmouth Dante Project
NJ New Brunswick (Rutgers Univ): Medieval and Early Modern Data Bank
NJ New Brunswick (Rutgers): Lexicon Iconographicum Mythologiae Classicae
NJ Princeton (Inst Advanced Studies): Greek Inscriptions from Asia Minor
NJ Princeton (Univ): American Founding Fathers Project
NY Binghamton (SUNY): Italian Madrigal and Related Reportories: Indexes to Printed Collections, 1500-1600
NY Buffalo (SUNY): WNY-ARCH = Western New York Archaeology
NY Ithaca (Cornell Univ): Greek Inscriptions from Attica
NY Ithaca (Cornell Univ): Cornell Blake Concordance Texts
NY New York (Columbia Univ): Women in Religious Communities: Italy 500-1500
NY New York (Columbia Univ): Data Archive, Center for Social Sciences
NY New York (Columbia Univ): Data Base on Labor Unrest in Imperial Russia
NY New York (Columbia Univ): Great Dictionary of the Yiddish Language
NY New York (Columbia Univ): Buddhist Canon Project
NY New York (Jewish Theological Seminary): Talmud Text Databank
NY New York (NYU): The Verdi Archive
OH Cleveland (Cleveland State Univ): Century-of-Prose Corpus
PA Philadelphia (Drexel Univ): The Latin Writings of Milton
PA Philadelphia (Univ PA): CCAT = Center for Computer Analysis of Texts
PA Philadelphia (Univ PA): Language Analysis Project
RI Providence (Brown Univ): WWP = Women Writers Project
RI Providence (Brown Univ): Romanian Love Incantations
RI Providence (Brown) ??: Nelson Francis Brown Corpus
TX Dallas (Theol Seminary): Biblical Data Bank (CD-ROM)
TX Del Valle (??): Chol (Mayan) Dictionary Database
TX Edinburg (Pan American Univ): RGFA + Rio Grande Folklore Archive
UT Provo (Brigham Young Univ): HRC = Humanities Research Center
UT Salt Lake City (Church of Latter Day Saints): Genealogical Data
WI Madison (Univ WI): DOSL = Dictionary of the Old Spanish Language
WI Milwaukee (Marquette Univ): Works of Karl Rahner Project